group13model

group13

2023-03-16

Data Wrangling and Pre-processing

Removing Value

Clean and process datasets by removing missing data and outliers, then select specific rows based on certain criteria.

```
dataset13 <- read_csv("dataset13.csv")
newdataset<- na.omit(dataset13)
newdataset<- newdataset%>%
  arrange(desc(altitude_mean_meters))
newdataset<- newdataset[-c(1:4),]
newdataset<- newdataset%>%
  arrange(aroma)
newdataset<- newdataset[-1,]
str(newdataset)
```

Calculating Correlation

In order to prepare for subsequent improvement and selection of variables during modelling, we firstly calculated the correlation between every two numerical variables.

```
newdataset[,2:6]%>%
  cor()%>%
  kable(caption='\\label{tab:correlation} correlation between 5 numerical variables')%>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: correlation between 5 numerical variables

	aroma	flavor	acidity	category_two_defects	altitude_mean_meters
aroma	1.0000000	0.7253135	0.5907547	-0.1934092	0.1632542
flavor	0.7253135	1.0000000	0.7438336	-0.2477485	0.1476604
acidity	0.5907547	0.7438336	1.0000000	-0.1851076	0.1778057
category_two_defects	-0.1934092	-0.2477485	-0.1851076	1.0000000	-0.0025717
altitude_mean_meters	0.1632542	0.1476604	0.1778057	-0.0025717	1.0000000

Table 1 shows the correlation between every two variables including aroma, flavor, acidity, category_two_defects and altitude_mean_meters. We can see that the correlation between aroma& flavor

(0.725) and the correlation between flavor&acidity (0.744) are both more than 0.7, which means these pairs have strong positive correlation. There is also a moderate correlation between aroma and acidity (0.591), while the correlation between other pairs are relatively weak.

Processing Non-numerical Data

For non-numerical data, including country_of_origin, Qualityclass and harvested(year), we set the country_of_origin and harvested(year) as factors. While as a qualitative variable, we converted Qualityclass into dummy variables, 'poor' to '0' and 'good' to '1'.

```
names(newdataset)
newdataset$country_of_origin<- as.factor(newdataset$country_of_origin)
newdataset$Qualityclass<- ifelse(newdataset$Qualityclass=='Poor',0,1)
newdataset$harvested <- as.factor(newdataset$harvested)
```

Formal Data Analysis

We used GLM to fit a logistic regression model with Qualityclass as the binary response variable, and country_of_origin, aroma, flavor, acidity, category_two_defects, altitude_mean_meaters and harvested as the explanatory variables. A summary of the model and a graph showing the points estimate for the log-odds with their corresponding 95% confidence interval are obtained as results.

Basic GLM

```
mod.cafe <- glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects +
                  altitude_mean_meters + harvested, data = newdataset, family = binomial(link = "logit")
print(summary(mod.cafe)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested,
    family = binomial(link = "logit"), data = newdataset)
```

```
tidy(mod.cafe)
```

```
# A tibble: 46 x 5
   term                            estimate std.error statistic  p.value
   <chr>                              <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -145.       11.6   -12.4     2.15e-35
 2 country_of_originBurundi           1.91      4.93    0.387   6.98e- 1
 3 country_of_originChina             0.500     1.08    0.462   6.44e- 1
 4 country_of_originColombia          1.82      0.564   3.22    1.29e- 3
 5 country_of_originCosta Rica        0.290     0.763   0.380   7.04e- 1
 6 country_of_originCote d?Ivoire   -12.1    6523.     -0.00186 9.99e- 1
 7 country_of_originEcuador          -1.43      1.50   -0.954   3.40e- 1
 8 country_of_originEl Salvador       0.541     0.958   0.565   5.72e- 1
 9 country_of_originEthiopia         13.3     898.      0.0148  9.88e- 1
10 country_of_originGuatemala        -0.583     0.546  -1.07    2.85e- 1
# ... with 36 more rows
```

```
#AIC = 543
```

##Plot of distribution

```
p <- plot_model(mod.cafe, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.5)
p + theme(plot.margin = unit(c(1,1,1,1), "cm"),
          axis.text.x = element_text(margin = margin(t = 10)),
          axis.text.y = element_text(margin = margin(r = 10)))
```

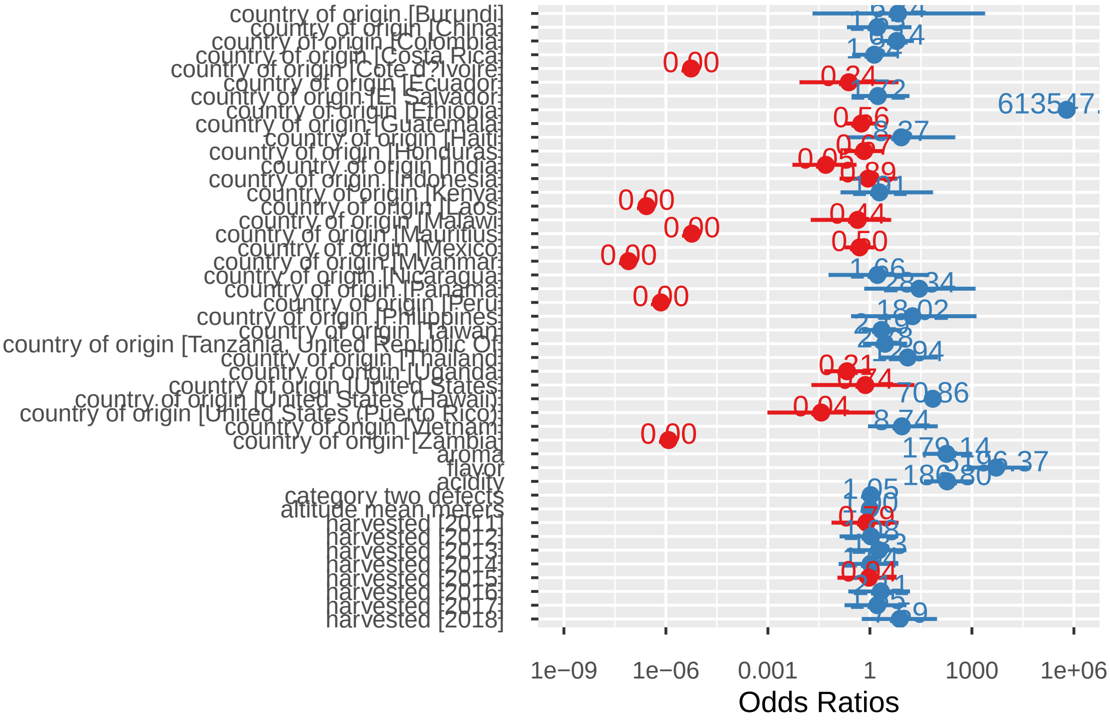

Figure 1: Odds of various factors influencing the quality of coffee(basic GLM model)

In the results we can see that aroma, flavor and acidity has coefficients of 5.19, 8.56, 5,23 separately, indicating comparatively strong positive influence on cafe quality, whilst category_two_defects and altitude_mean_meters do not appear to have much impact. For country_of_origin and harvested, different countries and vintages have different degrees of influence on the quality of coffee. For example, Ethiopia has a strong capacity to produce good coffee, as it has a coefficient more than ten(13.33) , Panama and Hawaii have lower coefficients(3.34 and 4.26), but still can be a good places to make coffee. However, there are countries with coefficients below -10,like Cote d'Ivoire,Laos,Mauritius, Myanmar,Peru and Zambia, which shows that they are likely to produce poorer cafe. In addition, only the harvested of 2018 shows a little positive impact on cafe quality(2.03), while other variables do not appear to be strongly influential.

GLM Stepwise

In the previous basic GLM we fitted a model with AIC of 543, wondering whether there is better regression to fit the data after selecting only the influencial variables, we then decided to use stepwise regression to improve our model.

```
# Fit a glm using stepwise regression with AIC as the criterion
model.step <- stepAIC(glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_de
```

```
# Print the selected model
print(summary(model.step)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters, family = binomial(link = "logit"),
    data = newdataset)
```

```
tidy(model.step)
```

```
# A tibble: 38 x 5
   term                            estimate std.error statistic  p.value
   <chr>                              <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -139.        11.0   -12.7     5.64e-37
 2 country_of_originBurundi           1.78       5.23    0.341   7.33e- 1
 3 country_of_originChina            -0.0540     1.01   -0.0535  9.57e- 1
 4 country_of_originColombia          1.59       0.525   3.03    2.45e- 3
 5 country_of_originCosta Rica        0.0645     0.714   0.0903  9.28e- 1
 6 country_of_originCote d?Ivoire   -12.1     6523.     -0.00185 9.99e- 1
 7 country_of_originEcuador          -1.38       1.48   -0.938   3.48e- 1
 8 country_of_originEl Salvador       0.468      0.942   0.497   6.19e- 1
 9 country_of_originEthiopia         13.5      886.      0.0152  9.88e- 1
10 country_of_originGuatemala        -0.776      0.497  -1.56    1.18e- 1
# ... with 28 more rows
```

```
#AIC = 537
p <- plot_model(model.step, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.50)
p + theme(plot.margin = unit(c(1,1,1,1), "cm"),
          axis.text.x = element_text(margin = margin(t = 10)),
          axis.text.y = element_text(margin = margin(r = 10)))
```

Using stepwise regression, we fitted a model with AIC of 537, which is relatively smaller than 543 in our first basic model, hence we can say that stepwise regression helped us to improve our model.

Adding Interaction Terms

Considering the possible interactions between the variables, based on the previously calculated correlations,we added some interaction terms(aroma:flavor,flavor:acidity,aroma:acidity)in order to improve our model.We summarized the results and graphically showed the log-odds and their corresponding 95% confidence intervals.

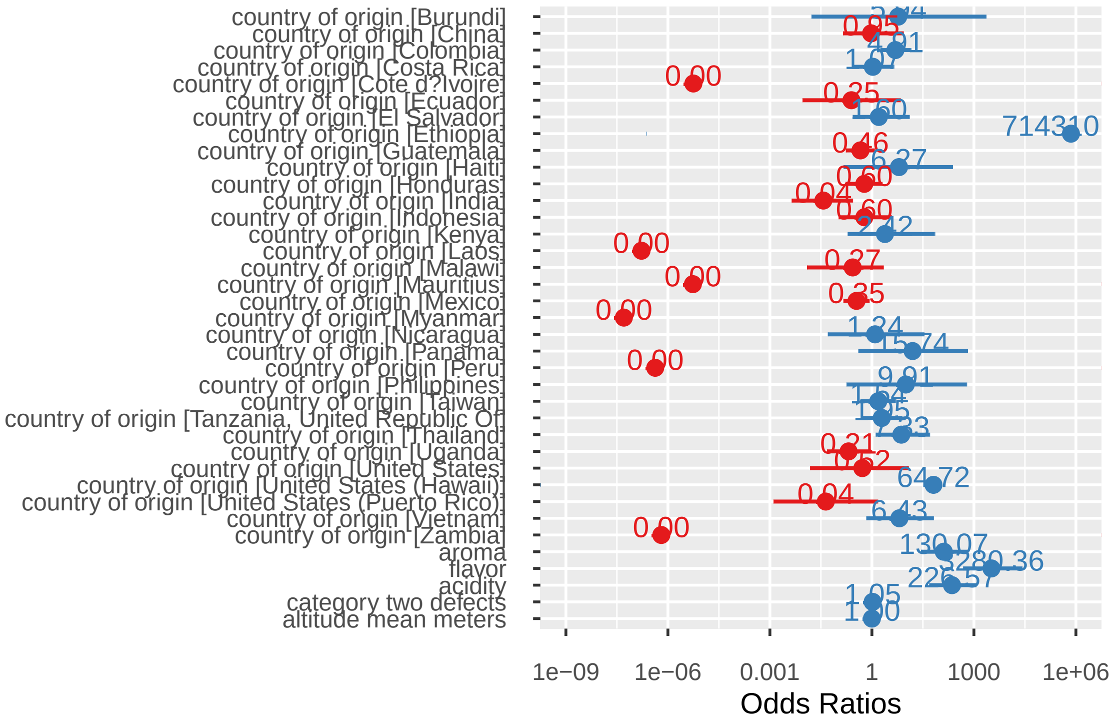

Figure 2: Odds of various factors influencing the quality of coffee(basic GLM model)

```
mod.cafe <- glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + al
                   family = binomial(link = "logit"))
print(summary(mod.cafe)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested +
    aroma:flavor + flavor:acidity + aroma:acidity, family = binomial(link = "logit"),
    data = newdataset)
```

```
tidy(mod.cafe)
```

```
# A tibble: 49 x 5
   term                            estimate std.error statistic  p.value
   <chr>                              <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -1185.       312.     -3.80    0.000144
 2 country_of_originBurundi           3.30        9.72    0.339   0.734
 3 country_of_originChina             0.498       1.08    0.462   0.644
 4 country_of_originColombia          1.78        0.582   3.06    0.00224
 5 country_of_originCosta Rica        0.0985      0.774   0.127   0.899
 6 country_of_originCote d?Ivoire   -10.4      6523.     -0.00159 0.999
 7 country_of_originEcuador          -1.51        1.51   -1.00    0.317
 8 country_of_originEl Salvador       0.543       0.968   0.561   0.575
 9 country_of_originEthiopia         14.6      617.       0.0236  0.981
10 country_of_originGuatemala        -0.591      0.557   -1.06    0.288
# ... with 39 more rows
```

```
#AIC = 539
```

```
p <- plot_model(mod.cafe, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.5)
p + theme(plot.margin = unit(c(1,1,1,1), "cm"),
          axis.text.x = element_text(margin = margin(t = 10)),
          axis.text.y = element_text(margin = margin(r = 10)))
```

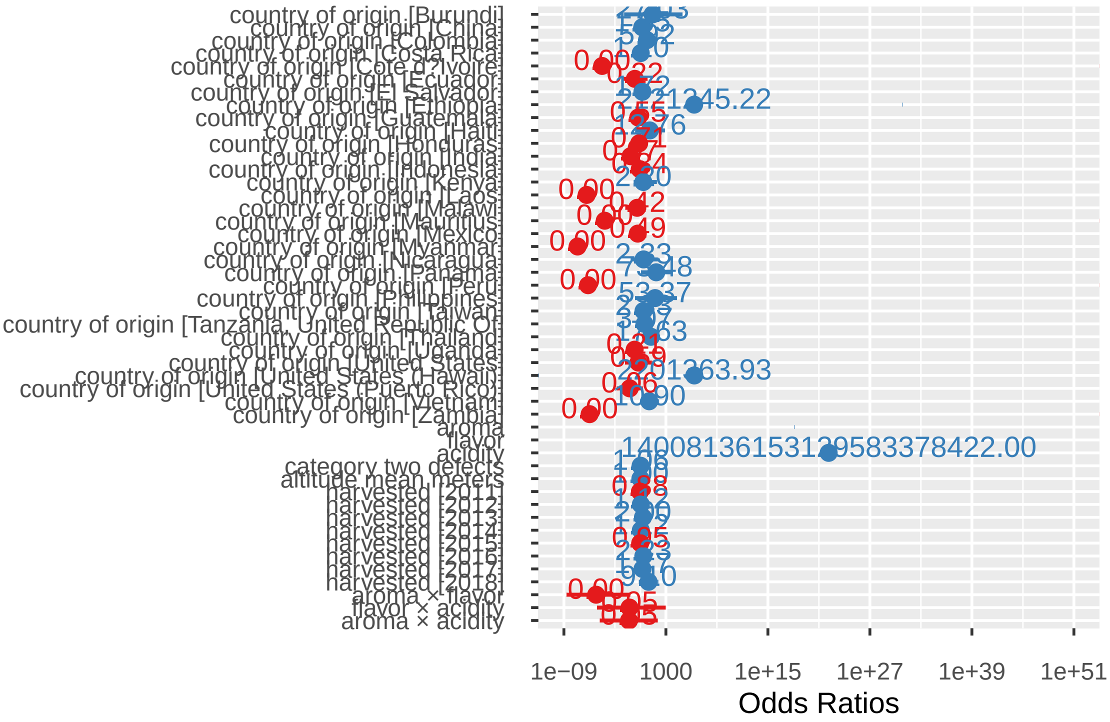

Figure 3: Odds of various factors influencing the quality of coffee(model with interaction terms)

In the results we can see the coefficients of aroma, flavor and acidity themselves are significantly positive, while the coefficients of all our possible interaction terms are negative,which shows that these three variables may moderate each other.After adding interaction terms, we can find that the AIC of the model decreases compared to the basic model, thus we can assume that the addition of the interaction terms improved our model.

GLM Stepwise After Adding Interaction Terms

In order to further improve our model, we fitted the GLM model with interaction terms using the method of stepwise regression with AIC as the criterion.

```
model.step <- stepAIC(glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_de
```

```
print(summary(model.step)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + aroma:flavor,
    family = binomial(link = "logit"), data = newdataset)
```

```
tidy(model.step)
```

```
# A tibble: 39 x 5
   term                            estimate std.error statistic  p.value
   <chr>                              <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -856.       237.     -3.61    0.000305
 2 country_of_originBurundi           2.91       8.89    0.328   0.743
 3 country_of_originChina            -0.0960     1.00   -0.0957  0.924
 4 country_of_originColombia          1.55       0.533   2.92    0.00354
 5 country_of_originCosta Rica       -0.0852     0.711  -0.120   0.905
 6 country_of_originCote d?Ivoire  -11.3     6523.      -0.00173 0.999
 7 country_of_originEcuador          -1.43       1.48   -0.966   0.334
 8 country_of_originEl Salvador       0.575      0.943   0.610   0.542
 9 country_of_originEthiopia         13.5      892.      0.0151  0.988
10 country_of_originGuatemala        -0.766      0.503  -1.52    0.128
# ... with 29 more rows
```

```
#AIC = 532
p <- plot_model(model.step, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.50)
p + theme(plot.margin = unit(c(1,1,1,1), "cm"),
          axis.text.x = element_text(margin = margin(t = 10)),
          axis.text.y = element_text(margin = margin(r = 10)))
```

We can see from the results that the AIC decreased to the lowest among these four models we fitted. AS AIC balances simplicity and accuracy when evaluating models, we can say that after adding an interaction term and doing the stepwise regression, our fourth model is the best model. Also, the last model has the lowest BIC=720, while the other three are 766,721,776 separately, which further demonstrates the superiority of our model.

```
levels(newdataset$country_of_origin)
```

```
 [1] "Brazil"                         "Burundi"
 [3] "China"                          "Colombia"
 [5] "Costa Rica"                     "Cote d?Ivoire"
 [7] "Ecuador"                        "El Salvador"
 [9] "Ethiopia"                       "Guatemala"
[11] "Haiti"                          "Honduras"
[13] "India"                          "Indonesia"
[15] "Kenya"                          "Laos"
[17] "Malawi"                         "Mauritius"
[19] "Mexico"                         "Myanmar"
[21] "Nicaragua"                      "Panama"
```

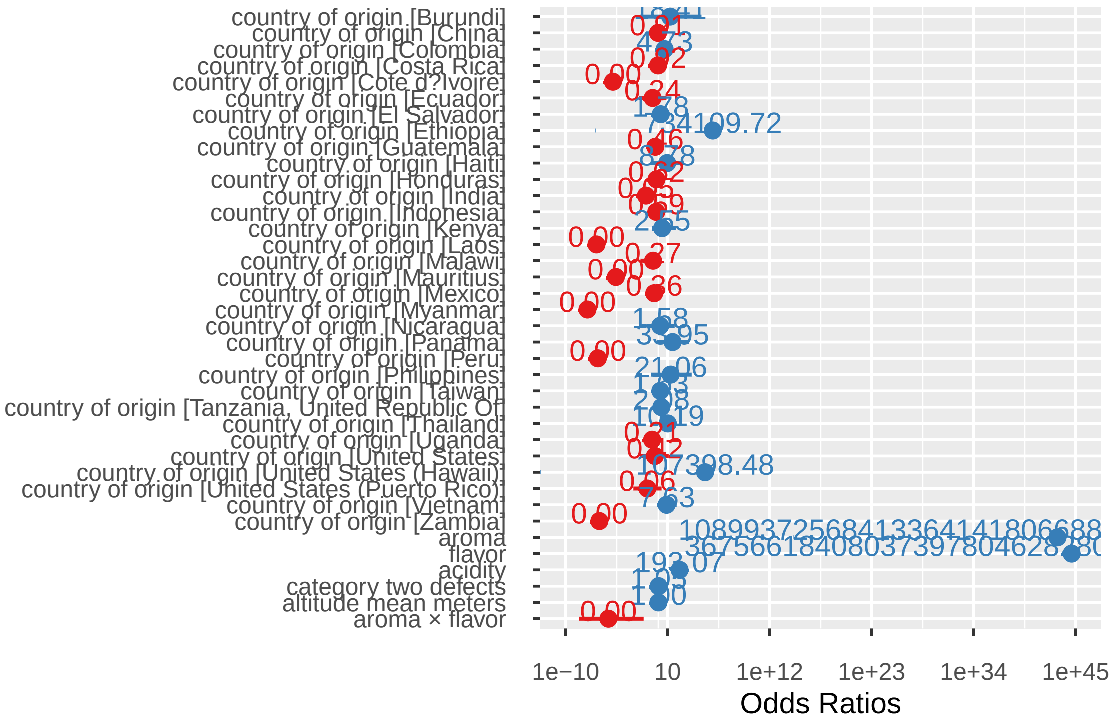

Figure 4: Odds of various factors influencing the quality of coffee(stepwise regression with interaction terms)

```
[23] "Peru"                         "Philippines"
[25] "Taiwan"                       "Tanzania, United Republic Of"
[27] "Thailand"                     "Uganda"
[29] "United States"                "United States (Hawaii)"
[31] "United States (Puerto Rico)"  "Vietnam"
[33] "Zambia"
```

checking assumptions

Residuals Plots for each variables

```
res <- resid(mod.cafe)
par(mfrow=c(3,2))
plot(newdataset$aroma,res,xlab='aroma')
abline(0,0)
plot(newdataset$flavor,res,xlab='flavor')
abline(0,0)
plot(newdataset$acidity,res,xlab='acidity')
abline(0,0)
plot(newdataset$category_two_defects,res,xlab='category two defects')
```

```
abline(0,0)
plot(newdataset$altitude_mean_meters,res,xlab='altitude mean meters')
abline(0,0)
plot(newdataset$harvested,res,xlab='harvested',ylab='res')
abline(0,0)
```

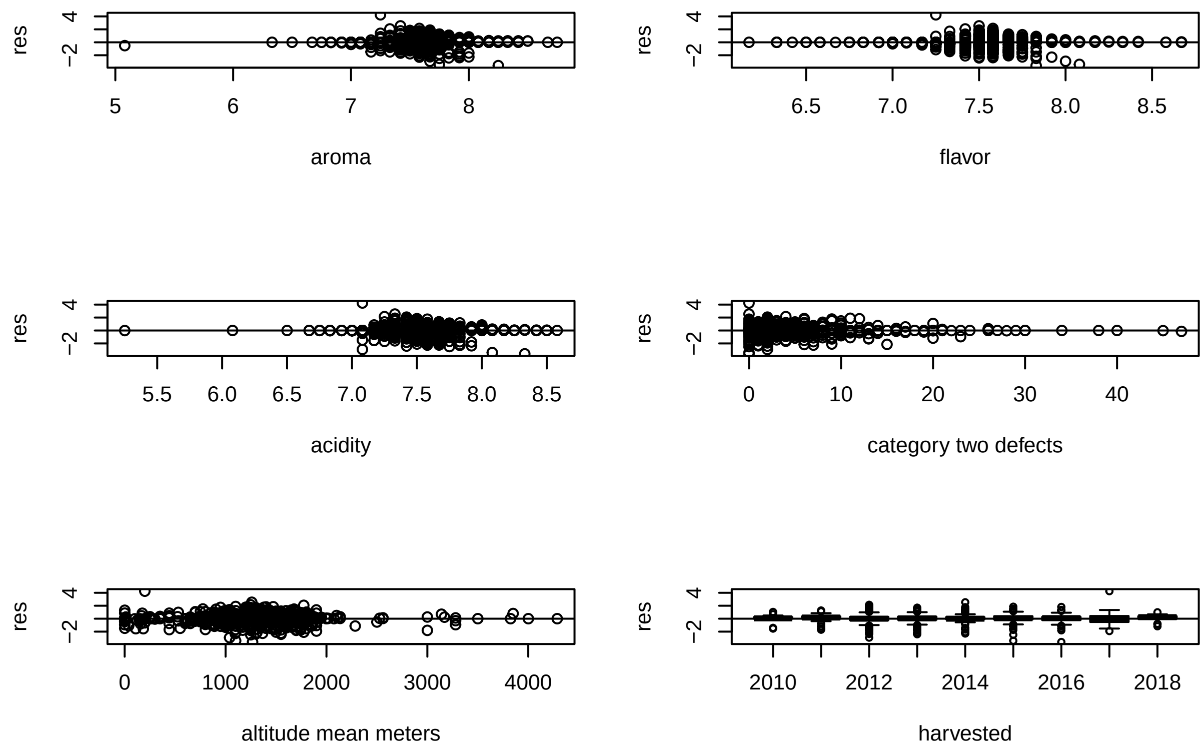

Figure 5: residuals against each variables

We see that there is an even spread of the residuals above and below the zero line for each variables,although there are a very few outlier points, overrall their spread on the graphs are acceptable, hence our assumption that the residuals have mean zero appears valid.

Density Plot

```
plot(density(res),xlab='residuals',title='')
```

In the graph we can see that the residuals are normally distributed with the mean 0, therefore the assumption is valid.

The remaining assumptions hold naturally at the time of our modelling, thus our model appears valid.

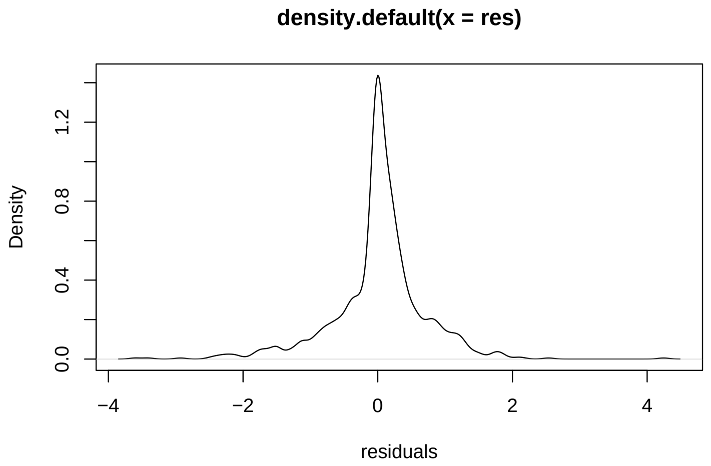

Figure 6: density plot of residuals

Conclusion

After data cleaning and processing of non-numerical data, we fitted the data to a regression model to observe the effect of each variable in the dataset on coffee quality, and we continued to improve the model by stepwise regression and adding possible interaction terms, resulting in the model with the smallest AIC value and therefore the most profile accurate model4. Looking at the summaries and graphs of model4, we can pick out the factors that have the greatest impact: aroma and flavor are very positively influencing on the quality of coffee, with coefficients of 99.1 and 102.62. The influence of origin varies very much, for example, Ethiopia and Hawaii contributes pretty much, for their coefficients are both more than ten(13.51 and 11.58). While Cote d'Ivoire,Laos,Mauritius, Myanmar,Peru and Zambia has less than -10 coefficients,making relatively strong negative effects. However, the many remaining origins do not seem to have any impact on the quality of the coffee.